

Name:

Student ID#:

Statistical Pattern Recognition (CE-725)
Department of Computer Engineering
Quiz #1 (Overview & Introductory Materials) - Spring 2011

1. (5 points) What is the difference between following terms: "Data Reduction", "Dimension Reduction", "Feature Selection", "Feature Extraction" and "Sampling"?

Data reduction implies on reduction the size of samples, but, dimension reduction implies on reduction the size of features. Moreover, data reduction works in pattern space while dimension reduction works in the feature space. However, data reduction also reduces the complexity of the feature space. Reducing the feature space complexity led to the reduction of model (classifier) complexity.

Sampling is a method of data reduction, and, feature selection and feature extraction are methods for dimension reduction. Feature selection methods, select a subset of features, but feature extraction methods create new features, based on existing ones.

2. (5 points) Suppose we have the following data set. All variables, including the labels, are binary.

X1	X2	X3	Y
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	1
1	0	1	0
1	1	0	1
1	1	1	0

Suppose we can't store all 3 features in the memory and hence we have to remove one of the features. Which of the 3 features, if any, do you suggest for removal? Why? Why not others?

If we remove either X1 or X3, the remaining features will be inconsistent, i.e. for example (X2,X3)=(0,0) may yield to either Y=0 or Y=1. But it does not happen for (X1,X3) pair, So we may successfully remove X2.

3. Bonus (5 points) Explain the role of feature dimensionality in the problem of model overfitting?

The more the data dimensions, the more parameters to learn and hence more risk of overfitting and lack of generalization.